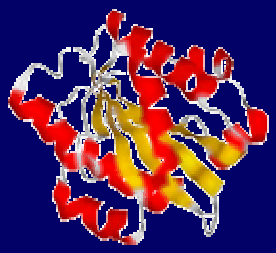
Protein Structure Predictions Using PROSPECT



Ying Xu
Protein Informatics Group
Life Sciences Division
Oak Ridge National Laboratory

R&D Goals and Our Strategies

Development of Computational Techniques for Reliable and Efficient Predictions of Protein Tertiary Structures

- Protein tertiary structure prediction by protein threading
- Improvement of threading accuracy through fully utilizing available experimental structural data
- Assessing prediction reliability by comparing general structural features with known structures

PROSPECT for Protein Structure Prediction

- PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit) is a threading-based prediction system developed at ORNL
- PROSPECT employs
  - a standard set of "energy" functions to measure quality of threading alignments
  - a unique divide-and-conquer algorithm to rigorously find globally optimal threading alignment
  - a unique capability for assessing prediction reliability

target sequence
MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE
template library

PROSPECT Threading Problem and Algorithm

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

how preferable to put two particular residues nearby: E_p

alignment gap penalty: E_g

how well a residue fits a structural environment: E_s

total score: E_p + E_s + E_g

- Solves the threading problem, considering residue-residue contacts
  - Residue-residue contacts effective for fold recognition but not much for alignments
- Achieves computing efficiency by considering only short-range contacts and utilizing the fact proteins have low topological complexity

Prediction Reliability Assessment

- Examine feature space of threading alignments:

  (singleton score, pair contact scores, secondary structure score, hydrophobic moment score, ...) *versus* true/false fold recognition

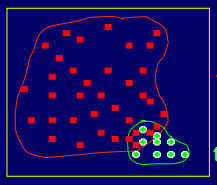
  -2000, -500, -35, -90, ..., true

  -1000, -201, -11, -500, ..., false

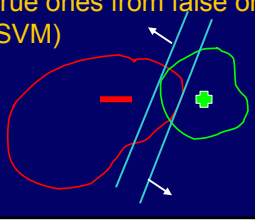
  -5020, -900, -20, -75, ..., true

  -1050, -185, -18, -320, ..., false

  .....



- Separate true ones from false ones using support vector machine (SVM)

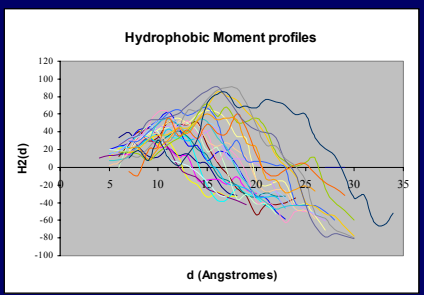


Prediction Reliability Assessment

- Each feature has somewhat different distributions in the true and false predictions
- E.g., hydrophobic moments (Hydrophobic moments of protein structures: spatially profiling the distribution, David Silverman, *PNAS* 2001 98: 4996-5001) is quite useful in distinguishing true from false threading predictions

target	template	threading rank
T0102	1bo9a	25
T0107	1mfm	315

In those two cases, all higher ranked models have bad hydrophobic moment profiles



PROSPCT Performance

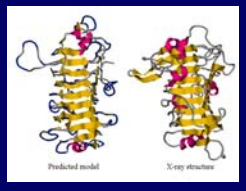
- On our "standard" training/testing data set with 311 query-template pairs with < 25% sequence identities
  - alignment accuracy on training set (174 pairs) improved from 72.0% in FY2000 to 79.5% now
  - alignment accuracy on testing set (137 pairs) improved from 70.6% in FY2000 to 79.5% now, using our improved energy functions
  - fold recognition accuracy has improved significantly using a combination of new information and methods

	Top 1 Model	Top 5 Models
FY2000	59.1%	72.6%
Now	79%	87%

PROSPECT Performance

- Prediction examples from CASP4

Among 32 targets, PROSPECT recognized correct folds for 25, of which 12 are not recognizable using PSI-BLAST



target 100:
no detectable homologue in PDB
5A for significant portion of the backbone

Table IV: Predictors ranked by total score *

Rank by T_i	Rank by Q_i	Predictor name	Group name	code	number of submissions
1	5	David Baker	baker	354	34
2	1	Alexey G. Murzin	Murzin	384	16
3	3	Michael J E Sternberg	Sternberg	126	24
4	6	Kevin Karplus	SAM-T2K	94	29
5	7	Leszek Rychlewski	BioInfo.PL	31	30
6	13	Ying Xu	ORNL-PROSPECT	88	33
7	13	Daniel Fischer	Cafasp-consensus	359	30
7	15	Richard A. Friesner	Friesner	414	31
9	11	Adam Godzik	Godzik	197	27
10	8	Burkhard Rost	rost	77	24
10	17	Nickolai Alexandrov	Walts-Wondrous-Wizards	44	31
12	17	Lawrence Kelley	Sternberg-3DPSSM	132	30
13	4	Andrei N Lupas	SBfold	381	17
13	10	Barry Honig	Honig-Barry	42	23
15	2	Tom L. Blundell	blundell-tl	95	10
15	30	Eckart Bindewald	BinToHes	255	34
17	9	Andrei Lomize	Lomize-Andrei	2	21
18	21	Leszek Rychlewski	FFAS	395	27
18	25	David Jones	Jones	23	30
18	27	Daniel Fischer	Fischer-Daniel	357	31

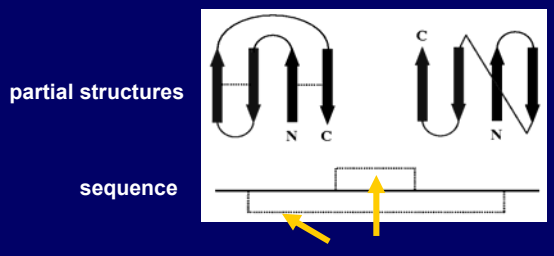
Sippl *et al*, "Assessment of CASP4 Fold Recognition Category", *PROTEIN* special CASP4 issue (in press), 2002.

Data-Constrained Threading

- Some structural information may be available before whole structure is solved
  - disulfide bonds
  - active sites
  - residues identified buried/exposed
  - (partial) secondary structure
  - partial NMR data
  - inter-residual distances by cross-linking and mass spec
- These data can provide highly useful constraints on threading prediction

Data-Constrained Threading

- in search for optimal threading alignment, consider only alignments that are consistent with specified constraints

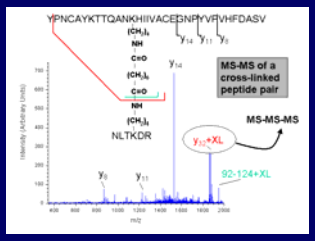


two disulfide bonds

PROSPECT provides a rigorous framework for dealing with threading constraints

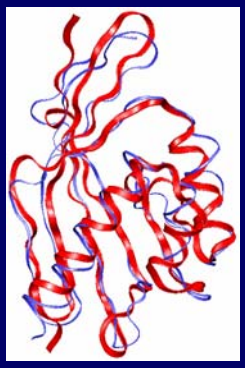
Getting Threading Constraints

- Residue-residue contacts identified through analyses of mass spectral data of cross-linked proteins



- Inter-residue distances from NMR/NOE experiments
  - limited NOE data can be collected/assigned in early phase of NMR experiments
  - for some proteins, only sparse NOE data are obtainable

Structure Calculation using Threading and Early NMR data



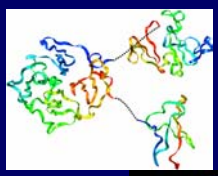
Blue: predicted model

Red: NMR structure

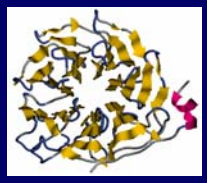
- target: *ketosteroid isomerase* KSI (Wu et al, 1997)
- early data: secondary structure/backbone NOEs
- template: 1ounA; *C_alpha-RMSD* 1.97A and seq. identity 9% with KSI
- threading and post-threading modeling with early NMR data
  - 2.73A RMSD for all heavy backbone atoms

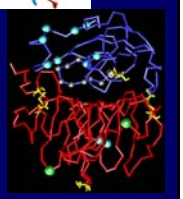
Applications of PROSPECT

- Victronectin: a three-domain protein
  - our collaborator has provided various structural data form her experiments, including disulfide bonds, active sites, heparin binding sites, cleavage sites, ...
  - we have done data-constrained threading/docking to its structure prediction



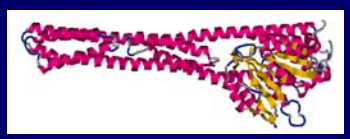
- COP-1: developmental regulator
  - we have done data-constrained threading using provided experimental structural data from our collaborator





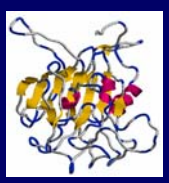
Applications of PROSPECT

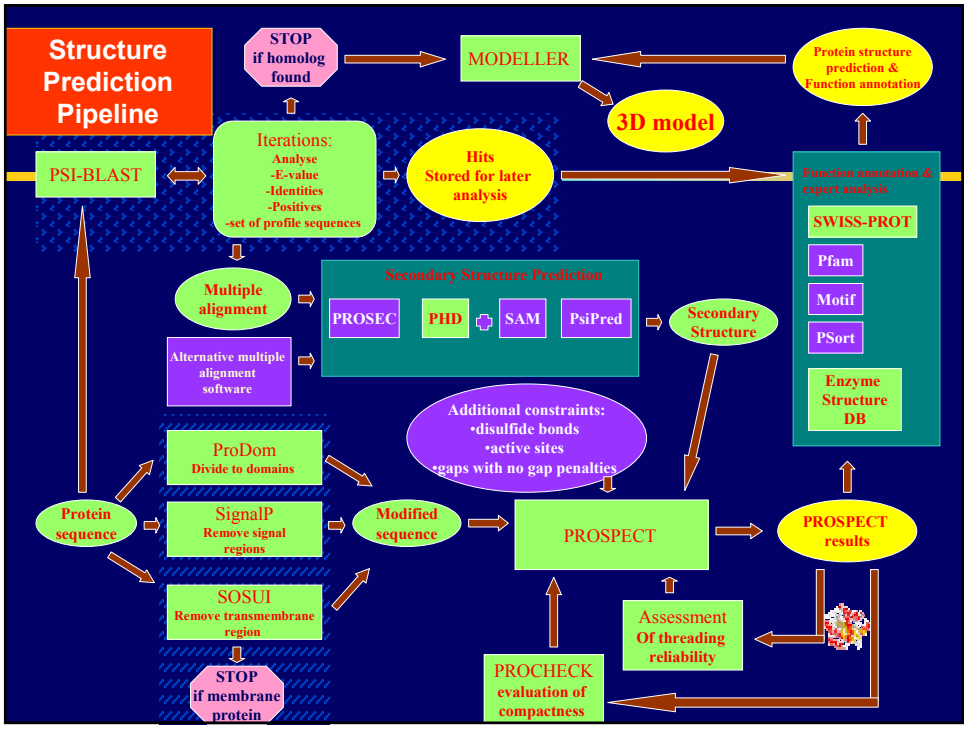
- Prediction of hypothetical proteins in *Shewanella oneidensis* MR-1, identified from microarray data (J. Zhou's lab)
- Example: ORF 3403



1. High prediction reliability based on our assessment
2. Template used: an interferon induced gluanylate binding protein
3. Made some functional inference

- Examples: ORF 1964 and ORF 3208





Future Directions

- More “accurate” threading energy functions
- Faster threading algorithms
- More effective methods for prediction reliability assessment

- Large-scale applications to microbial genomes
- From structures to complex structures
- From structures to functions
- ....

Acknowledgements

- PROSPECT R&D:
  - Dong Xu
  - Dongsup Kim
  - Juntao Gao
  - Victor Olman
  - Oakley Crawford
- Applications:
  - Christal Secrest
  - Yu Chen
  - Jane Razumavskaya
  - Can Ozen
  - Kyle Ellrott

DOE/OBER for Funding

- Prediction pipeline:
  - Manesh Shah
  - Serguei Passovets
  - Li Wang
  - Phil LoCascio
- Collaborators:
  - Michael Summers (UMBC)
  - Engin Serpersu (UTK)
  - Frank Larimer (ORNL)
  - Ed Uberbacher (ORNL)
  - Cynthia Peterson (UTK)
  - Jizhong Zhou (ORNL)
  - Greg Hurst (ORNL)

Structure Calculation by Combining Threading and NMR

- NMR-constrained threading: *using partial NMR data as threading constraints*
  - improve threading accuracy
  - expand scope of threading to structural analogs
- threading-supported NMR method: *adding information from threading to NMR calculation*
  - reduce the amount of NMR data needed for NMR structure calculation
- Predicted structure can help NMR data assignments
  - through matching NMR spectra with calculated spectra